主动采样与标注估计技术研究及应用

◎ 吴伟宁 著

科学出版社

主动采样与标注估计技术研究及应用

吴伟宁 著

科学出版社

北京

内 容 简 介

主动学习的理论及其应用是机器学习研究领域中一个富有生命力和备受关注的研究分支，现已成为解决实际问题的重要方法之一. 本书集中介绍主动学习方法中的一些典型的样本选择方法和标注估计策略，并给出主动学习在应用中的统一框架. 本书通过研究大量丰富的文献资料和科研成果，回顾主动学习的过去，分析主动学习的研究现状，继而对主动学习的未来进行充分展望.

本书可供高等院校计算机、自动化、电子工程等专业的高年级本科生、研究生、教师及相关领域的研究人员与工程技术人员参考.

图书在版编目(CIP)数据

主动采样与标注估计技术研究及应用/吴伟宁著. —北京: 科学出版社, 2017. 6

ISBN 978-7-03-053324-1

Ⅰ. ①主… Ⅱ. ①吴… Ⅲ. ①噪声-采样-研究 Ⅳ. ①O422.8

中国版本图书馆 CIP 数据核字 (2017) 第 130606 号

责任编辑: 李静科 / 责任校对: 贾伟娟
责任印制: 张 伟 / 封面设计: 陈 敬

科学出版社出版
北京东黄城根北街 16 号
邮政编码: 100717
http://www.sciencep.com

北京京华虎彩印刷有限公司印刷
科学出版社发行 各地新华书店经销
*
2017 年 6 月第 一 版 开本: 720 × 1000 B5
2017 年10月第二次印刷 印张: 6 1/4 插页: 4
字数: 89 000

定价: 48.00 元

(如有印装质量问题, 我社负责调换)

前　言

机器学习是人工智能领域的一个重要分支. 近年来, 随着互联网技术和数据存储技术的迅速发展, 利用机器学习算法从所收集的数据中获取其隐藏的知识, 并利用这些知识改善程序在实际任务中的性能, 成为备受关注的方向之一. 目前, 随着所收集数据规模的日趋增大, 半监督学习和主动学习等技术侧重于同时利用标注数据和未标注数据来进行知识发现, 这一做法在实际任务中得到了广泛的使用. 与仅利用标注数据的学习方法相比, 这类技术试图利用未标注数据来辅助学习过程. 其中, 主动学习技术从模拟人类自身的学习行为出发, 从大量未标注数据中选择最有帮助的部分样本加入学习过程, 充分利用有限的人类标注者经验来改善学习系统自身性能.

本书主要总结了主动学习近年来的主要工作, 包括主动学习的理论工作及其在实际任务中的应用, 涵盖了国内外关于主动学习许多具有代表性的最新研究成果; 详细介绍了主动学习近年来涌现出的理论和应用方面的现有工作, 并从代价的角度给出了主动学习的基本工作框架; 根据所收集数据条件的不同, 从样本选择和添加标注两个方面, 介绍了主动学习在实际任务中的策略, 包括如何从噪声数据中选取样本和添加标注信息, 以及如何在大规模数据中降低学习的时间花销等. 全书内容丰富, 注重理论与实际的结合, 着重介绍了噪声和大规模数据条件下如何选择样本和添加标注, 即如何在非实验室理想数据环境下降低主动学习过程中的标注代价和时间花销. 全书分为 5 章: 第 1 章绪论, 主要介绍了主动学习的基本思想、研究现状等; 第 2 章重点阐述了主动学习中的加权样本选择方法; 第 3 章阐述了基于分布优化的主动样本选择过程; 第 4 章介绍了多个标注者同时提供标注信息的条件下如何为所选择的样本添加正确标注; 第 5 章介绍了如何在大规模数据条件下快速选择样本, 降低学习过程中的时间花销.

本书内容源于 863 计划项目 (项目编号：2007AA01Z171), 国家自然科学基金项目 (项目编号：61171185). 本书得到了国家自然科学基金 (项目编号：61502117) 和黑龙江省科学基金 (项目编号：QC2016084) 的资助. 本书的研究工作得到了郭茂祖教授和刘扬副教授的指导和帮助, 并获得了黄少滨教授的关心和支持, 作者对他们致以深切的谢意.

由于作者的水平有限, 书中难免存在不妥之处, 恳请广大读者批评指正.

吴伟宁

2016 年 1 月

目　录

第1章 绪 论

学习能力是人类具有的极其重要的特征之一. 在人的学习过程中, 学生向老师提问是快速掌握知识的有效方法. 机器学习在人工智能研究中占据着非常重要的地位, 是人工智能领域的核心内容之一, 其目的是使计算机程序具有类似于人类的学习能力, 从观测到的数据中获取经验或者知识. 其中, 主动学习模拟了人类的学习过程, 通过向人类专家 "提问" 这种拟人化方式来学习新的知识.

1.1 主动学习的背景

随着数据处理技术的飞速发展, 机器学习已经被广泛应用于各种工业领域和社会生活中, 它在人类的科研、生活、沟通与交流等各个方面发挥着重要的作用. 例如: 计算机视觉中的遥感图像分析[1,2]、医学图像分析[3,4]、基于内容的图像分类与视频检索[5–9], 以及文本挖掘[13–15]、语音识别[16,17]、生物信息挖掘[18] 等. 机器学习的研究目标是使计算机具有通过单一或者一组数据, 获取周围数据环境信息的能力, 即让计算机实现人类的学习功能, 感知、识别和理解客观世界的场景和行为, 从而帮助人们准确快速地从浩瀚的各类数据中搜索和获取重要的信息. 在此过程中, 数据对象的分析与理解是描述数据内容, 获取相关知识必不可少的重要组成部分, 而对数据进行分类, 从而识别和判断给定数据对象中是否包含某种知识则是正确分析和理解数据语义内容的重要问题之一.

目前, 在模式识别领域, 解决数据对象分类问题的一般做法是: 预先收集一组数据, 利用收集数据中包含的语义信息作为先验知识, 判断其余数据对象的所属类别. 其分类过程可以描述如下: 分别提取已知和未知数据的特征, 构建训练集和测试集, 继而, 在训练集上利用机器学习或模式识别技术建立算法或模型, 利用训练好的模型对测试集中的数据进行分类, 得到待识别数据的类别信息. 由于该分类过程需要通过人工方式为同类别数据添加标注等监督信息, 目的是建立学习模型所需要的训练集,

获得较高精度的分类模型, 因而这一做法也称作监督学习.

在大多数已有监督学习系统中, 训练集是通过随机选择的方式构造的, 因而需要很高的人工标注代价. 随着互联网技术的迅猛发展和逐渐普及, 所收集数据的类别和规模呈现爆炸式增长, 因此, 减少训练集的人工标注代价和时间消耗成为该领域中一个亟须解决的重要问题. 上述监督学习过程面临的现实问题包括以下三个方面.

(1) 虽然互联网技术的发展使得学习系统可以通过关键字搜索在短时间内获得大量同类别的数据, 但是这些数据缺少精确标注. 如果对这些数据不加以选择和甄别, 直接对其添加标注信息会耗费大量时间和精力; 而让标注者自行选择和标注一部分数据, 这样构造的训练集又带有很强的偏好信息和个人倾向性, 即标注者所选择的数据并不一定是学习系统最需要的. 而且, 已有研究结果表明[10], 标注者自行选择的数据和学习系统收集的数据往往存在明显的不同和差异. 因此, 该做法不能很好地利用人工标注资源, 甚至浪费了有限的人力资源, 限制了标注者知识向学习系统的迁移.

(2) 在所收集数据包含信息是否有利于训练的问题上, 标注者与学习系统的判断存在着较大的差异. 因此, 学习系统本身应当具备判断未标注数据中信息含量的能力, 并知晓哪些数据对自身模型训练是最有利的, 进而选择这部分数据并提交标注者添加标注, 建立训练集. 这种通过学习系统和标注者进行交互来建立训练集的做法对充分利用标注者的监督信息、降低标注代价和克服人工选择数据中的偏好信息具有重要的意义和很高的实用价值. 另外, 当标注者之间存在不同的偏好差异时, 学习系统应当能够从观测到的标注信息中捕捉标注者的倾向性信息, 并将最有帮助的数据提交给最合适的标注者, 这种能力对于获取有利于模型训练的监督信息无疑是十分有利的. 最后, 学习系统应当以最快的速度向标注者提交查询请求, 减少选择所消耗的时间, 这对于减少构建训练集需要的时间代价具有重要作用.

(3) 使用互联网技术收集得到的数据建立训练集, 学习系统面临以下两个问题: 第一, 通过互联网技术收集得到的数据集中, 各个类别对应的数据数量之间往往存在较大差异 (例如: 在图像分类问题中, PASCAL VOC 图像库[11], MIRFLICKR 图像库[12] 等), 其中, 某些类别包含的数

据数量远远小于其他类别, 这种类别不平衡问题直接影响了所训练模型的性能和准确度; 第二, 互联网环境是不断变化的, 而学习系统掌握的先验知识却是一次性从收集数据中得到的, 学习系统无法得知外部环境的变化情况, 也不能动态获取和补充先验知识, 这极大地限制了学习系统的适应性. 为了解决这两个问题, 学习系统应当具备良好的适应能力, 在不同类别和外部环境下, 根据不同类别, 选择最有利的数据构建训练集, 扩展模型的先验知识. 这对于在相同标注代价下, 提升模型的准确度具有重要的意义.

因此, 针对这些监督学习系统面临的实际问题, 在传统监督学习基础上, 涌现出了一批新的算法和模型 —— 目的是通过增加学习系统与外部环境和标注者的交互能力, 减少人工标注代价, 克服人工收集数据的倾向性, 尽可能将标注者知识迁移到学习系统中, 从而提高学习系统识别概念的能力. 鉴于这类机器学习方法众多, 无法一一展现, 本书仅介绍主动学习技术, 着重介绍如何使用主动学习技术解决分类问题, 主要包括有效的样本选择和标注者选择的方法, 目的是使学习系统以最少的人工标注代价和时间消耗选择对自身训练最有帮助的数据集.

1.2 主动学习的技术特点

与监督学习被动构建训练集的方法不同, 主动学习模拟了人类的学习过程, 即将在训练集中已标注数据上学习得到的知识作为先验信息, 利用该先验知识对测试分布中未标注数据包含的信息进行判断, 选择对模型训练最有利的数据进行标注, 以达到减少分类模型训练过程所需标注代价的目的. 以图像分类任务为例, 在图 1-1 中, 本书给出了主动学习构建分类系统的训练过程和预测过程. 从图 1-1 可以看出, 主动学习增加了样本选择和查询标注信息这两个环节, 其训练样本是通过设计有效的样本选择方法从未标注样本分布中选择得到的, 样本对应的标注是通过向标注者进行查询所得. 这两个步骤增加了机器学习系统与外部环境和标注者的交互能力, 提高了学习系统的适应性. 使用主动学习对分类问题建立模型, 可以充分利用珍贵的标注者资源, 降低学习过程中必要的标注代价, 这一做法的优势如下:

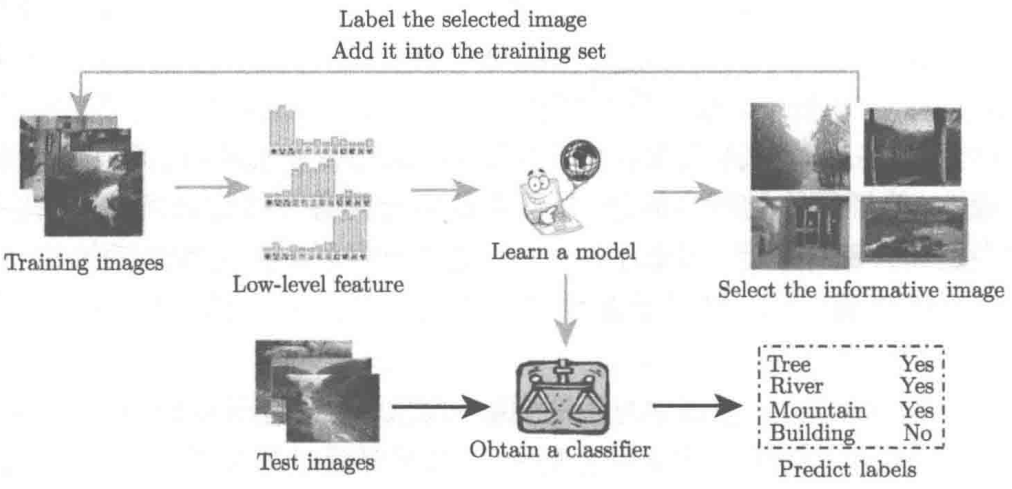

图 1-1　基于主动学习的图像分类过程示意图 (详见文后彩图)

训练过程使用蓝色箭头标出, 预测过程使用黑色箭头标出, 所选择图像使用红色方框标出

(1) 主动学习通过设计合适的样本选择方法从未标注数据中选择部分样本标注和建立训练集, 让分类系统可以根据需要自行选择样本, 克服了标注者自行选择和标注样本带来的个人偏好和倾向性. 同时, 由于系统选择了对当前模型训练最有利的样本加入学习过程, 当标注代价相同时, 与随机选择样本加入训练过程的监督学习相比, 主动学习可以获得更大程度上的分类性能提升.

(2) 主动学习通过向标注者提交查询请求的方式获取训练数据标注信息, 在这个迭代过程中, 主动学习可以从已经观测到的标注信息中, 判断标注者的喜好, 并在接下来的查询过程中利用掌握的偏好信息, 将所选样本交给最恰当的标注者. 这一学习系统与标注者进行交互的做法可以更加充分地利用有限的标注者资源, 更好地将标注者知识大规模迁移到学习系统当中.

(3) 主动学习通过判断未标注数据的信息含量来选择训练样本, 这一做法增加了学习系统探查外部环境的能力, 使学习系统可以根据不同类别和测试分布建立训练集. 当类别不同或测试分布随时间、地点不同而逐渐变化时, 主动学习可以更好地调整训练分布, 减少不必要的标注代价. 当标注代价相同时, 从不同的外部环境中, 分类模型可以更好地获取先验知识, 提升模型性能.

因此, 本书首先总结了已有主动学习技术, 随后介绍了不同的数据环境下, 如何发掘有效的样本选择策略和代价度量函数. 在标注代价相同的条件下, 准确设计上述两者对减少系统的学习代价具有重要的意义, 也是提升分类模型性能的重要手段之一.

1.3 主动学习的研究现状

机器学习作为人工智能的一个重要研究方向, 一直受到计算机科学家的关注. 当前, 机器学习面临的现实情况是: 未标注样本数目众多, 易于获得; 标注样本数量稀少, 难以获得. 一些研究表明, 对于训练样本的精确标注不但需要该领域中大量的标注者参与, 并且标注样本花费的时间是其获取时间的 10 倍以上[20], 与之形成对比的是, 未标注样本却简单易得. 这种现实使传统机器学习方法无法得以有效应用, 原因在于监督学习需要大量的标注样本对分类器进行迭代训练, 否则根据 PAC 学习理论, 算法的泛化性能无法有效提高. 无监督学习虽然直接利用未标注样本, 但是算法的精度不能使人满意. 在这种情况下, 半监督学习 (Semi-supervised Learning) 和主动学习 (Active Learning) 算法就应运而生并迅速发展, 成为解决上述问题的重要技术. 虽然两者都利用未标注样本和标注样本共同构建高精确度的分类器, 降低人类标注者的工作量. 但不同的是, 主动学习算法模拟了人的学习过程, 选择标注部分样本加入训练集, 迭代提高分类器的泛化性能, 因此近年来被大量地应用于信息检索、图像和语音识别、文本分类和自然语言处理等领域. 2009 年, Tomanek 和 Olsson[21] 的一项调查显示, 90.7% 的研究者认为主动学习在他们的项目应用中是有效的. 而另一份调查证明 Google, CiteSeer, IBM, Microsoft 和 Siemens 等大型公司也都在项目中使用主动学习来提高性能[19]. 2010 年, PASCAL 举办了主动学习方法竞赛, 竞赛包含 6 个不同应用领域, 目的是鼓励参赛者开发优秀的主动学习方法[22].

主动学习最初是由耶鲁大学的 Angluin 教授在*Queries and concept learning* 一文中提出的[23]. 与以往学习方法的不同点是该文中使用了未标注来辅助分类器的训练过程, 其方法是选择并标注部分未标注样本, 然后放入标注样本集训练分类器, 使用分类器再次选择未标注样本. 这种

有选择地扩大有标注样本集和循环训练的方法使分类器获得了更强的泛化能力. 此后, 由于主动学习的适用性广泛和高效利用人类标注者资源等一系列特点, 使得这种学习方式得以迅速发展, 并成为机器学习领域最重要的方向之一. 例如, 卡内基–梅隆大学、斯坦福大学的机器学习实验室都将主动学习的算法理论以及实际应用, 特别是无标注样本选择方法的设计列为研究重点; 一些机器学习、数据挖掘的重要学术会议也都收录主动学习的文章并将其列为重要专题进行讨论[24].

1.3.1　主动学习过程

简单说来, 主动学习对应的工作过程是一个迭代训练分类器的过程, 该过程由以下两个部分组成[25].

- 学习引擎 (Learning Engine, LE): 学习引擎的工作过程是在标注样本集合上进行循环训练, 当达到一定精度后输出. 这一过程类似于监督学习中的分类器训练过程.
- 采样引擎 (Sampling Engine, SE): 采样引擎是主动学习不同于其他学习方法的部分. 其任务是使用不同的样本选择方法选取未标注样本, 将其交给标注者以获取标注信息, 并将标注后的样本加入标注样本集, 以供分类器进行循环训练. 这一过程试图在标注代价最少的条件下获取最有助于分类器训练的标注样本集合.

主动学习的迭代过程可以被描述为: 在标注样本集上训练分类器; 使用分类器对未标注样本进行类别判断; 根据判断结果, 使用采样引擎选择部分未标注样本提交标注者添加标注信息; 将标注后的样本加入训练集用于分类器的下一次训练. 终止条件是标注代价或者分类器的泛化精度达到一定标准为止. 为了说明主动学习的工作过程, 图 1-2 给出了主动学习过程的伪代码描述.

主动学习中采样引擎的核心是样本选择方法, 它决定了主动学习的实际应用效果. 因此, 本书主要介绍了主动学习中的样本选择方法, 侧重介绍样本选择过程和标注添加技术. 其中, 样本选择方法的目的是设计合适的样本信息含量度量标准, 也就是所选样本被加入训练集后对分类器泛化能力的影响程度. 根据无标注样本的选择方式不同, 该标准可以是一个设定好的函数, 也可以是一个固定的阈值. 标注添加技术的目的

是在不同的标注者环境下准确地为所选择的无标注样本添加合适的标注信息, 也就是从人类标注者提供的标注信息中估计所选样本对应的正确标注信息. 通过以上两种技术, 主动学习在分类训练过程和样本标注代价之间进行选择, 试图在花费代价最小的条件下, 达到学习系统增益最大的目的.

主动学习过程的伪代码描述

输入：标注样本集L, 无标注样本集U, 学习引擎LE, 采样引擎SE
输出：学习引擎LE

```
BeginFor i=1,2,…,N
1.    Train(LE,L);              //在标注样本集上学习分类模型
2.    T=Test(LE,U);             //使用该分类模型预测未标注样本的类别信息
3.    S=Select(SE,U/T);         //使用采样引擎选择未标注样本
4.    Label(S);                 //将所选样本提交标注者获取标注信息
5.    L=L+S;                    //将标注样本加入标注样本集
6.    U=U-S;                    //从无标注样本集中删除所选样本
BeginFor
```

图 1-2 主动学习过程的伪代码描述

1.3.2 主动学习分类

根据样本选择方法选取未标注样本的方式不同, 可以将主动学习分为以下三种：成员查询综合 (Membership Query Synthesis)、基于流 (Stream-based) 的主动学习和基于池 (Pool-based) 的主动学习[19]. 为了便于叙述, 本书使用 (x, y) 表示已标注样本及其对应标注信息, 使用 $\bar{x}$ 表示未标注样本.

成员查询综合是最早被提出的使用查询进行学习的思想[23], 即假定学习系统对周围环境具有一定控制能力, 可以向人类标注者提问. 算法通过提问的方式确定某些样本的标注和学习未知概念. 该方法的缺点是将所有未标注样本都交给人类标注者进行标注, 而不考虑样本的实际分布情况[26].

针对这一缺陷, 研究人员提出了一系列样本选择算法对该方法进行改进. 当 $\bar{x}$ 大量易得时, Cohn 提出标注 $p(x, y)$ 超过某一阈值的样本[27].

Seung 等提出在 (x, y) 上分别训练参数为 θ_1 和 θ_2 的两个模型, 选择这两个模型预测不一致的 $\bar{x}$ 进行标注[28]. 这类做法也称为基于流的采样策略, 其采样过程是将落在版本空间 (Version Space) 中的所有 $\bar{x}$ 按照顺序逐个依次进行标注[29], 并广泛应用于词类标注[30]、信息检索[31]、入侵检测[32] 和信息提取[33] 等实际问题.

虽然基于流的样本选择在一定程度上解决了直接查询方法的问题, 但是这种样本选择方法往往需要设定一个固定阈值来衡量样本的信息含量, 因此缺乏对不同学习问题的普适性. 具体应用问题不同, 设定的阈值也不同. 更重要的是, 算法需要逐个将 $\bar{x}$ 的信息含量与标准阈值进行比较, 故无法掌握 $\bar{x}$ 的实际分布, 也无法得知 $\bar{x}$ 之间的差异.

为了解决上述问题, Lewis 提出将 $\bar{x}$ 组成一个无标注样本 "池", 主动学习从这个集合中选择样本, 即基于池的样本选择策略[34]. 与基于流的样本选择策略相比, 算法维护一个固定分布的由大量 $\bar{x}$ 组成的 "样本池". 主动样本选择方法逐一计算 $\bar{x}$ 的信息含量并比较, 选择信息含量高的 $\bar{x}$ 进行标注. 由于基于池的样本选择策略继承了前面两种方法的优点, 克服了它们的不足, 因而它成为当前研究最充分、应用最广泛的样本选择策略, 在文本分类[35–37]、信息提取[38]、图像检索[39,40]、视频检索[41,42] 和癌症检测[43] 等领域都有具体的应用.

1.3.3　主动学习的理论分析

主动学习的目的是减少训练所需的标注代价, 因此, 在主动学习理论研究中, 备受关注的是算法对样本复杂度 (Sample Complexity) 的降低程度. 相对于主动学习的大量应用研究工作而言, 该方面的理论研究依然有很多开放性问题. 特别是目前主动学习已有研究成果大多针对特定条件或模型, 尚缺乏一般性结论. 根据主动学习中理论研究针对的不同问题, 将该方向成果划分为 "可达" (Realizable) 和 "不可达" (Non-realizable) 两种情形, 并分别加以阐述.

可达类主动学习是指假设类 (Hypothesis Class) 中存在可以完美划分数据的假设. 对于可达情形, 其理论研究是主动学习理论研究中被关注时间较早, 相对较为充分的一种类型. 大多数该方面的理论工作证明: 相对于监督学习而言, 主动学习可以有效降低样本复杂度. 与监督学习相比,

主动学习可以产生“指数级”的样本复杂度改善[44−49]. 例如, Cohn 等[27]证明在标准 PAC 模型下, 均匀分布的样本空间中, 获得一个最大错误率为 ε 的分类假设. 监督学习需要的样本复杂度是 $O(1/\varepsilon)$, 而主动学习使用二分搜索获得该分类假设所需要的样本复杂度为 $O(\log 1/\varepsilon)$. Freund 等[49] 进一步提出, 在贝叶斯条件下, 获得一个泛化误差小于 ε 的分类假设, 基于版本空间缩减的主动学习算法的样本复杂度为 $O(d\log(1/\varepsilon))$ (d 表示当前空间中 VC 维的维度). 而相同条件下, 监督学习的样本复杂度为 $O(d/\varepsilon)$. 针对该结论, Gilad-Bachrach 使用核方法进一步限制版本空间的大小, 获得了更高的性能[50]. Balcan 等[47] 证明样本均匀分布时, 可达类主动学习的样本复杂度是 $O\left(\varepsilon^{-2(1+\lambda)}\right)$, 其中 λ 表示噪声参数.

但是, 由于存在噪声数据或者假设类学习能力有限等因素, 实际应用中的大多数问题属于“不可达”情形, 即假设类中不存在对数据进行完美划分的假设. 针对不可达类主动学习, 人们获得了丰富的研究成果. 这部分研究成果可以根据是否假定噪声模型划分为以下两种.

在不假定噪声模型的条件下, 一些研究成果[51−53] 表明, 主动学习的样本复杂度下界与被动学习的样本复杂度上界相当, 这意味着, 主动学习并不能起到实质性的改善作用. 因此, 基于 PAC 框架的不可达主动学习算法的特点是严格地限制采样次数, 从而达到降低样本复杂度的目的. 例如, Balcan 等[51] 提出基于 PAC 框架的不可达主动学习(Agnostic Active Learning, A^2) 算法, 证明了样本复杂度边界是 $O(\ln 1/\varepsilon)$. 该结果表明只要样本是从一个固定分布中选择的, 主动学习就比监督学习算法具有更大的优势. Steve Hanneke[52] 通过定义不一致系数, 进一步限制样本复杂度的上界. 在此基础上, Dasgupta 等[53] 提出一种更有效的样本选择方法, 针对不同的样本分布和模型类别, 更精确地对当前假设进行限制, 达到了更少的标注代价. 这些主动学习算法类似于精确枚举空间中的所有假设, 计算复杂度高, 所以很难直接应用于实际. 而且, 算法的理论分析结论往往建立在样本均匀分布或近似均匀分布的条件下, 或者对假设空间有严格要求. 在算法具体实现中, 这些方法局限于优化简单的 0-1 损失函数, 很难扩展到复杂监督模型或其他对象函数[54]. 值得注意的是, Wang 和 Zhou 使用全类扩张的 α- 扩张定义, 显示出主动学习算法在数据存在多视图时降低样本复杂度的有效性[55], 同时, Wang 和

Zhou 也将半监督技术与多视图主动学习相结合, 并获得算法性能的进一步提高.

在假定噪声模型的条件下, 现在一般考虑 Tsybakov 噪声模型, 又可以分为有界 (Bounded) 和无界 (Unbounded) 两种情形. 有界情形相对简单, 一些研究表明[45,56,57], 主动学习可以产生指数级的样本复杂度改善. 例如, Kaariainen[58] 证明了噪声率为 η 时主动学习算法的样本复杂度为 $\Omega\left(\eta^2/\varepsilon^2\right)$. 无界情形则相对复杂, 也更接近于大多数真实问题, 一些研究工作[45,56,59,60] 表明主动学习仅能获得多项式级的样本复杂度改善, 并不能起到实质性提高. 例如, Cavallanti 等[59] 进一步提出, 当标注噪声满足线性条件时, 主动学习算法的样本复杂度是 $O\left(\varepsilon^{-2(3+\lambda)(1+\lambda)(2+\lambda)}\right)$. Castro 和 Nowak[60] 提出了单视图主动学习算法的样本复杂度的一般形式, 即获得一个分类错误率小于 ε 的分类假设, 样本复杂度至少为 $\Omega\left(\varepsilon^{-\rho}\right), \rho \in (0,2)$. 在分类边界和分布高阶平滑至无限平滑的假设条件下, Wang[61] 基于放松的 Tsybakov 噪声模型 (Approximately Tsybakov Model) 获得了指数级改善. Wang 和 Zhou[62] 首次发现, 当数据存在多视图时, 在无界 Tsybakov 噪声模型条件下, 主动学习可以达到指数级提升. 该工作具有很大的启发意义, 说明主动学习算法的进一步理论研究和算法设计都必须考虑一些具体的数据条件, 否则, 一般通用角度无法获得样本复杂度的指数级提升, 即不能获得实质性改善.

1.4　主动样本选择方法概述

一般来讲, 分类任务是对观察到的数据中包含实际物体、场景或行为等实体的所属类别做出有意义的判定, 该问题可以形式化地描述为: 在无标注样本集 I 上最大化样本判定类别 C 对应的条件概率, 即

$$C = \arg\max p\left(C|I,w\right) \tag{1-1}$$

其中, 参数 w 是分类模型经由主动学习系统迭代训练所得. 在每一轮迭代中, 样本选择过程根据 $p\left(C|I,w\right)$ 值计算未标注样本包含的信息含量, 并依据信息含量的高低选择最有助于分类模型训练的未标注样本, 提交标注者查询标注信息.